Минимизации некоторых робастных сумм параметризованных функций

З.М. Шибзухов[1,2], М.А. Казаков[1], Д.П. Димитриченко[1]

[1]Институт прикладной математики и автоматизации КБНЦ РАН, ул. Шортанова 89а, Нальчик, КБР, Россия, 360000
[2]Московский педагогический государственный университет, пр. Вернадского, 88, Москва, Россия, 119435

Аннотация. Рассматривается робастный подход к построению алгоритмов машинного обучения, основанный на минимизации робастных конечных сумм параметризованных функций. Он основывается на применении конечных робастных дифференцируемых агрегирующих функций суммирования, которые являются устойчивыми по отношению к выбросам.

1. Введение

Значительную часть задач машинного обучения можно поставить как задачу минимизации конечных сумм параметризованных функций:

$$\mathcal{Q}(\mathbf{w}) = \sum_{k=1}^{N} v_k \ell_k(\mathbf{w}),$$

где $\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})$ – неотрицательные базисные функции, $\mathbf{w}$ – вектор неизвестных параметров, $\mathbf{w} \in \mathbf{W} \subseteq \mathbb{R}^m$, $v_1, \ldots, v_N \geqslant 0$ – неотрицательные числовые веса. Чаще всего $v_k = \text{const}$, например, 1 (арифметическая сумма) или $1/N$ (среднее арифметическое). Оптимальный набор параметров $\mathbf{w}^*$ минимизирует целевую функцию $\mathcal{Q}$:

$$\mathcal{Q}(\mathbf{w}^*) = \min_{\mathbf{w} \in \mathbf{W}} \mathcal{Q}(\mathbf{w}).$$

Однако, если распределение значений базисных функций содержит выбросы, как последствие выбросов в обучающих данных, то минимизация $\mathcal{Q}(\mathbf{w})$, как правило приводит к искажению $\mathbf{w}^*$. Это происходит из-за того, что арифметическая сумма и среднее арифметическое не являются устойчивыми по отношению к выбросам.

Один из путей преодоления этой проблемы состоит в применении робастных агрегирующих функций для вычисления суммы или среднего. Таким образом появились определения для функции Q:

$$\mathcal{Q}(\mathbf{w}) = \operatorname*{med}_{k=1,\ldots N} \ell_k(\mathbf{w})$$

для робастных оценок среднего и

$$\mathcal{Q}(\mathbf{w}) = \sum_{k=1}^{N-p} \ell_{(k)}(\mathbf{w})$$

для робастных оценок суммы. Здесь $z_{(1)}, \ldots, z_{(N)}$ обозначает последовательность чисел, которая получается в результате упорядочивания по возрастанию исходной последовательности $z_1, \ldots, z_N$. Например, для построения робастной регрессии с $\ell_k(\mathbf{w}) = (f(\mathbf{x}_k, \mathbf{w}) - y_k)^2$ были предложены метод LMedS (Least Median Squares) и метод LTS (Least Treamed Squares) [1, 2]. Миинимизация таких оценок для конечных сумм или средних позволяет находить адекватные оценки для $\mathbf{w}^*$ при наличии выбросов в данных (вплоть до 50%). Однако алгоритмы минимизации включают комбинаторную составляющую в виде поиска $\mathbf{w}^*$по подвыбркам, т.к. они не являются дифференцируемыми. Это снижает масштабируемость таких алгоритмов и их применение для обучения нейронных сетей и в задачах с большими данными.

2. Минимизация M-средних от параметризованных функций

Для случая с медианой проблему можно преодолеть, используя M-средние [4], которые являются дифференцируемыми и, в определенном смысле, приближением медианы:

$$\mathsf{M}_\rho\{z_1, \ldots, z_N\} = \arg\min_u \sum_{k=1}^{N} \rho(z_k - u),$$

где ρ – неотрицательная строго выпуклая функция, $\rho(0) = 0$. Если ρ – дважды дифференцируемая, то$\mathsf{M}_\rho\{z_1, \ldots, z_N\}$ имеет все частные производные:

$$\frac{\partial \mathsf{M}_\rho}{\partial z_k} = \frac{\rho''(z_k)}{\rho''(z_1) + \cdots + \rho''(z_N)}.$$

Показано, что для "приближения" медианы, например, можно использовать следующие функции:

- $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$;
- $\rho_\varepsilon(r) = |r| - \varepsilon \ln(\varepsilon + |r|) - \varepsilon \ln \varepsilon$.

Такие M-средние M_{ρ_α}являются дифференцируемыми и при достаточно малых α являются робастными, так что они являются устойчными по отношению к выбросам (в некоторых случаях вплоть до 50%).

Для приближения α-квантиля можно использовать функцию

$$\rho_{\varepsilon,\alpha}(r) = \begin{cases} (1-\alpha)\rho_\varepsilon(r), & \text{если } r < 0 \\ \alpha\rho_\varepsilon(r), & \text{если } r \geqslant 0. \end{cases} \tag{1}$$

Алгоритм поиска $\mathbf{w}^*$ представляет собой процедуру итеративно перевзвешиванной минимизации эмпирического риска IR-ERM (Iteratively Reweighted Empirical Risk Minimization) [3]:

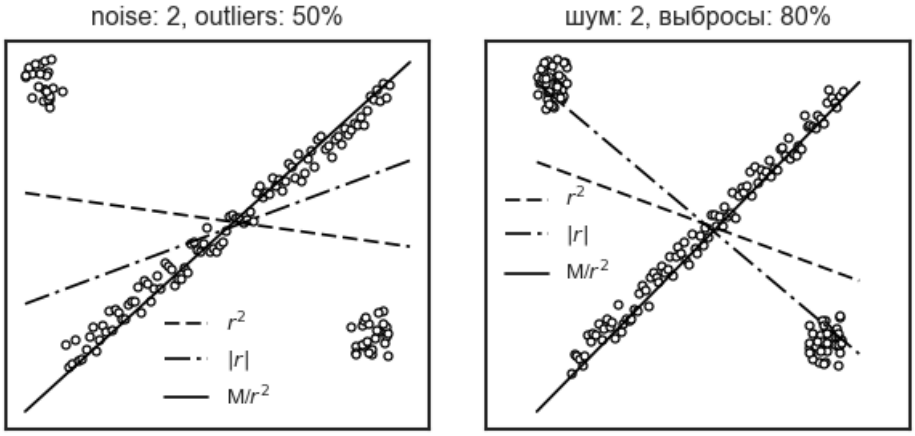

Рисунок 1. Примеры восстановления линейной регрессии пр наличии 50% и 80% выбросов от количества «нормальных» данных.

procedure IR-ERM($\mathbf{w}_0$)
  $t \leftarrow 0$
  repeat
    $z_1 = \ell_1(\mathbf{w}_t), \ldots, z_N = \ell_N(\mathbf{w}_t)$
    $\bar{z}_t \leftarrow \mathsf{M}\{z_1, \ldots, z_N\}$
    for $k = 1, \ldots, N$ **do**

$$v_k = \frac{\rho''(z_k - \bar{z}_t)}{\rho''(z_1 - \bar{z}_t) + \cdots + \rho''(z_N - \bar{z}_t)}$$

    end

$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w})$$

    $t \leftarrow t + 1$
  until $\{\bar{z}_t\}$ и $\{\mathbf{w}_t\}$ стабилизируются
end

Для демонстрации возможности принципа минимизации робастной оценки эмпирического риска и алгоритма IR-ERM приведем пример восстановления линейной регрессии при наличии большого числа выбросов в данных.

Данные представляют собой точки прямой с небольшой ошибкой, распределенной равномерно. Для восстановления применялся метод наименьших квадратов, метод минимизации абсолютных ошибок и метод минимизации робастной дифференцируемой оценки среднего при помощи M-среднего на основе функции

$$\rho_\alpha(r) = |u - z| - \alpha \ln(\alpha + |u - z|) + \alpha \ln \alpha,$$

где $\alpha = 0.001$. На Рис. 1 показаны результаты восстановления линейной регрессии. В обоих случаях метод минимизации робастной дифференцируемой оценки среднего позволил избежать влияния выбросов.

3. Минимизация робастных сумм функций

Рассмотрим ряд методов суммирования, которые могут быть устойчивыми по отношению к выбросам. Все M-средние, включая среднее арифметическое обладают свойством:

$$\frac{\partial \mathsf{M}}{\partial z_1} + \cdots + \frac{\partial \mathsf{M}}{\partial z_N} = 1.$$

Арифметическое суммирование же обладает следующим важным свойством:

$$\frac{\partial \mathsf{S}}{\partial z_1} + \cdots + \frac{\partial \mathsf{S}}{\partial z_N} = N.$$

Поэтому естественно, чтобы предлагаемые робастные методы суммирования также сохранили это свойство. Рассмотрим следующий метод суммированния.

Least Winsorized Sum (LWS) В этом методе перед суммированием все значения, которые больше, чем заданное пороговое значение u, заменяются на u, т.е.

$$\boxed{\mathsf{S}_u\{z_1, \ldots, z_N\} = \sum_{k=1}^{N} \frac{1}{2}(z_k + u - |z_k - u|)}.$$

Назовем его WS (Winsorised Sum). Он обладает следующим свойством: если u – среднее арифметическое от $z_1, \ldots, z_N$, то $\mathsf{S}_u\{z_1, \ldots, z_N\} = z_1 + \cdots + z_N$.

Обобщим метод LWS следующим образом. Пусть M_ρ – M-среднее на базе дважды дифференцируемой строго выпуклой функции ρ. Обозначим $\bar{z} = \mathsf{M}_\rho\{z_1, \ldots, z_N\}$. Определим

$$\boxed{\mathsf{S}_\rho\{z_1, \ldots, z_N\} = \sum_{k=1}^{N} \frac{1}{2}\big(z_k + \bar{z} - \rho(z_k - \bar{z})\big)}.$$

Вычислим частные производные:

$$\frac{\partial \mathsf{S}_\rho}{\partial z_k} = \frac{1}{2}\left(1 - \rho'(z_k - \bar{z})\right) + \frac{1}{2}\frac{\partial \mathsf{M}_\rho}{\partial z_k}\Big(N + \sum_{l=1}^{N} \rho'(z_l - \bar{z})\Big).$$

Т.к., по определению,

$$\sum_{k=1}^{N} \rho'(z_k - \bar{z}) = 0,$$

то

$$\frac{\partial \mathsf{S}_\rho}{\partial z_k} = \frac{1}{2}\left(1 - \rho'(z_k - \bar{z})\right) + \frac{N}{2}\frac{\partial \mathsf{M}_\rho}{\partial z_k}.$$

Поэтому

$$\sum_{k=1}^{N} \frac{\partial \mathsf{S}_\rho}{\partial z_k} = N.$$

Если

$$\lim_{|r|\to\infty} \rho(r)/|r| = 1,$$

то определенный здесь метод суммирования можно рассматривать как гладкий вариант WS.

Теперь можно рассмотреть следующую задачу минимизации целевой функции

$$\mathcal{Q}(\mathbf{w}) = \frac{1}{N}\mathsf{S}_\rho\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\}$$

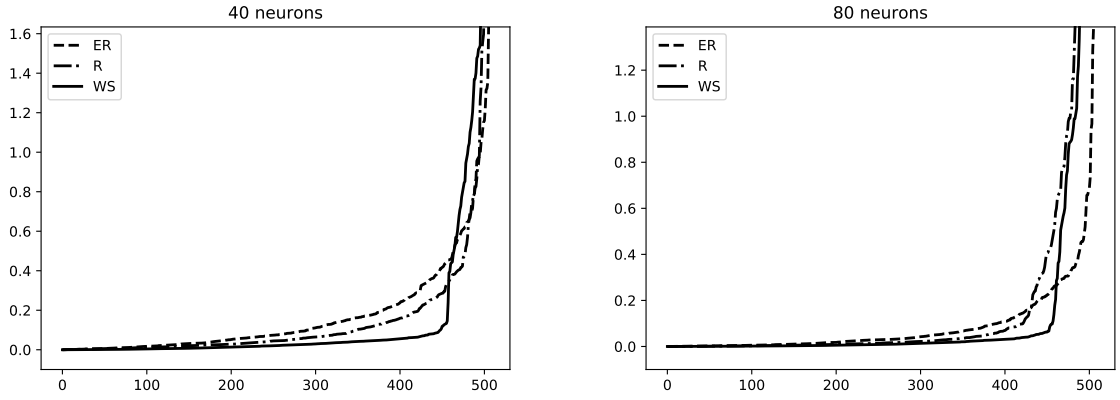

Рисунок 2. Распределение ошибок обученной НС с одним скрытым слоем, содержащим 40 и 80 нейронов, на наборе данны Boston.

для поиска оптимального набора параметров $\mathbf{w}^*$. Запишим градиент:

$$\operatorname{grad} \mathcal{Q}(\mathbf{w}) = \sum_{k=1}^{N} v_k(\mathbf{w}) \operatorname{grad} \ell_k(\mathbf{w}),$$

где

$$v_k(\mathbf{w}) = \frac{1}{2N}\left(1 - \rho'\big(\ell_k(\mathbf{w}) - \bar{z}(\mathbf{w})\big)\right) + \frac{1}{2}\frac{\partial \mathsf{M}_\rho}{\partial z_k},$$

а $\bar{z}(\mathbf{w}) = \mathsf{M}_\rho\{\ell_1(\mathbf{w}), \ldots, \ell_N(\mathbf{w})\}$. При этом, заметим, что

$$v_1(\mathbf{w}) + \cdots + v_N(\mathbf{w}) = 1.$$

Для численного расчета можно применить алгоритм IR-SLM (Iteratively Reweighted Sum of Losses Minimization) следующий вариант алгоритма IR-ERM:

procedure IR-SLM($\mathbf{w}_0$)
  $t \leftarrow 0$
  repeat
    $z_1 = \ell_1(\mathbf{w}_t), \ldots, z_N = \ell_N(\mathbf{w}_t)$
    $\bar{z}_t \leftarrow \mathsf{M}\{z_1, \ldots, z_N\}$
    for $k = 1, \ldots, N$ **do**
      $v_k = \frac{1}{2N}\left(1 - \rho'(z_k - \bar{z})\right) + \frac{1}{2}\frac{\partial \mathsf{M}_\rho}{\partial z_k}$
    end
    $\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_k \ell_k(\mathbf{w})$
    $t \leftarrow t + 1$
  until $\{\bar{z}_t\}$ и $\{\mathbf{w}_t\}$ стабилизируются
end

Для иллюстрации возможностей алгоритма IR-SLM рассмотрим задачу обучения нейронной сети на наборе данных Boston. Для обучения применялся метод обратного распространения ошибки, в котором минимизировалась средняя квадратичная ошибка (ER) и среднее значение от значений функции Хьюбера с малым значением параметра (0.001) для того, чтобы она представляла собой непрерывно дифференцируемую апроксимацию

функции модуля ошибки. Для обучения также применялся алгоритм IR-SLM, в котором минимизировалась робастная оценка суммы WS квадратов ошибки (WS) с функцией $\rho_{\varepsilon,\alpha}$ вида (1), где $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2} - \varepsilon$, $\alpha = 0.90$. На Рис. 2 представлены распределения абсолютных ошибок на всем наборе данных. Они наглядно демонстрируют, что при обучении НС при помощи алгоритма IR-SLM уменьшаются величины ошибки на более 80% данных.

4. Заключение

В настоящей работе предложен метод и алгоритмы минимизации робастных дифференцируемых оценок средних и сумм, которые являются потенциально устойчивыми к выбросам и ошибкам, которые могут приводить к смещению параметров обучаемых моделей.

5. Благодарности

Работа выполнена при поддержке гранта РФФИ 18–01–03381.

6. Литература

[1] Rousseeuw, P.J. Least Median of Squares Regression / P.J. Rousseeuw // Journal of the American Statistical Association. – 1984 – Vol. 79. - P. 871–880.

[2] Rousseeuw, P.J. Robust Regression and Outlier Detection / P.J. Rousseeuw, A.M. Leroy // NY: John Wiley and Sons, 1987.

[3] Shibzukhov, Z.M. On the Principle of Empirical Risk Minimization Based on Averaging Aggregation Functions / Z.M. Shibzukhov // Doklady Mathematics. – 2017 – Т. 96, № 2. – С. 494–497.

[4] Huber, P.J. Robust Statistics / P.J. Huber // NY: John Wiley and Sons – 1981.

[5] Vapnik, V. The Nature of Statistical Learning Theory / V. Vapnik // Information Science and Statistics. - Springer-Verlag, 2000.

[6] Хьюбер, П. Робастность в статистике / П. Хьюбер - М.: Мир, 1984.

[7] Mesiar R., Komornikova M., Kolesarova A., Calvo T. Aggregation functions: A revision. / H. Bustince, Herrera, J. Montero, editors // Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. Springer, Berlin, Heidelberg, 2008.

[8] Grabich M., Marichal J.-L., Pap E. Aggregation Functions. Series: Encyclopedia of Mathematics and its Applications. - Cambridge University Press, 2009. - Vol. 127.

[9] Beliakov G., Sola H., Calvo T. A Practical Guide to Averaging Functions. - Springer, 2016. - 329 p.

[10] Calvo T., Beliakov G. Aggregation functions based on penalties // Fuzzy Sets and Systems. - 2010. - Vol.161(10). - P. 1420-1436.

[11] Yohai, V.J. High breakdown-point and high effciency robust estimates for regression / V.J. Yohai // The Annals of Statistics. - 1987. - Vol. 15. - P. 642-656.

[12] Koenker R. Quantile Regression. NY: Campridge University Press, 2005.

[13] Newey W., Powell J. Asymmetric Least Squares Estimation and Testing // Econometrica. - 1987. - Vol. 55(4). - P. 819-847.

[14] Ma Y., Li L., Huang X., Wang S. Robust Support Vector Machine Using Least Median Loss Penalty // IFAC Proceedings Volumes (18th IFAC World Congress). - 2011. - Vol. 44(1). - P. 11208-11213.

[15] Shibzukhov Z.M. Correct Aggregate Operations with Algorithms // Pattern Recognition and Image Analysis. - 2014. - Vol. 24(3). - P. 377–382.

[16] Shibzukhov Z.M. Aggregation correct operations on algorithms // Doklady Mathematics. - 2015. - Vol. 91(3). - P. 391-393.

[17] Knigma D.P., Ba J. Adam: A Method for Stochastic Optimization. // CoRR. abs/1412.6980. - 2014. - URL: http://arxiv.org/abs/1412.6980}.

[18] Schmidt M., Le Roux N., Bach F. Minimizing Finite Sums with the Stochastic Average Gradient. // Mathematical Programming. - Springer-Verlag, 2016. - P. 1-30.

Minimization of robust sum of loss functions

Z. Shibzukhov[1,2], M. Kazakov[1], D. Dimitrichenko[1]

[1]Institute of Applied Mathematics and Automation KBSC RAS, Shortanova street 89a, Nalchik, Russia, 360000
[2]Moscow State Pedagogical University, Vernadskogo street 88, Moscow, Russia, 119435

Abstract. The problem of minimization of robust sums of parametric loss functions that arise when solving problems of classification and regression in the presence of a large number of outliers is considered. An iteratively reweighted method of minimizing empirical risk is proposed to search for parameters that minimize robust sum of parametric loss functions.

Keywords: machine learning, loss function, robust regression.